УДК 511.342.2

А. А. БОСОВ, Н. А. МУХИНА (ДИИТ)

МОДИФИЦИРОВАННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ В СТРУКТУРНОМ МОДЕЛИРОВАНИИ

Розглянуто задачу структурного моделювання за допомогою певним чином побудованого регресійного аналізу.

Рассмотрена задача структурного моделирования с помощью специальным образом организованного регрессионного анализа.

The problem of structural modelling with the help of specially organized regression analysis has been considered.

Проблема выбора лучших предикторов для заданного отклика остается актуальной при моделировании сложных систем. Прежде всего это связано с большим числом взаимосвязанных элементов, при этом взаимосвязи, как правило, заранее неизвестны. Если исследователь не имеет предварительной информации о порядке предикторов по их важности для предсказания отклика, то решение проблемы, как правило, сводится к регрессии отклика по всем возможным подмножествам показателей и выбору среди них наилучшего набора предикторов. Если при этом число показателей велико, то задача становится практически неразрешимой. Так, если необходимо y раскрыть как функцию от M переменных, то для выбора лучшей необходимо построить $2^M - 1$ моделей. Например, если число показателей равно $M = 20$, возникает необходимость построения $2^M - 1 = 1048575$ моделей, что становится труднообозримым при анализе. Одним из путей преодоления этих трудностей является пошаговая регрессия [1; 2]. В этом случае проблема состоит в том, какие именно показатели и в какой последовательности необходимо включать в структуру математической модели.

В работе [3] подробно рассматривается метод структурного моделирования, основанный на применении отношения толерантности τ. Этот метод обладает определенными преимуществами, так как при заданной доверительной вероятности позволяет определить структуру математической модели, не решая задачи параметрической идентификации.

Несмотря на отмеченные преимущества данного подхода, он существенно зависит от объема опытных данных, что требует применения и других методов для окончательного выбора предикторных переменных.

Исходная информация и ее преобразование

Пусть исследуемый объект определяется системой показателей $\Omega = \{x_1,\ x_2, ..., x_N\}$, информация о которых задана в виде матрицы экспериментальных значений.

Обозначим через y один из элементов множества Ω, а именно тот, для которого необходимо определить лучшие предикторы, и в дальнейшем будем называть его откликом модели. В этом случае матрицу экспериментальных значений можно представить в виде

$$[YX] = \begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{M1} \\ y_2 & x_{12} & x_{22} & \dots & x_{M2} \\ \dots & & \dots & \dots & \dots \\ y_N & x_{1N} & x_{2N} & \dots & x_{MN} \end{bmatrix},$$

где N – число периодов наблюдений или число однотипных объектов наблюдений.

Относительно числа наблюдений предположим, что $N > M$.

Для удобства столбцы матрицы $[YX]$ обозначим через $X_1, X_2, \dots, X_M$, причем в дальнейшем будем считать, что X_i, $i = 1,\ M$ – ортонормированные векторы, принадлежащие евклидовому пространству размерности N.

Необходимо отметить, что, как правило, это требование на практике не выполняется, поэтому матрицу экспериментальных значений $[YX]$ необходимо преобразовать в матрицу Z, столбцы которой являются ортонормированными векторами.

Для замкнутости изложения кратко опишем процедуру ортогонализации системы векторов.

Положим

$$Z_1 = X_1 / |X_1|,$$

где $|X_1| = \sqrt{\sum_{i=1}^{N} x_{1i}^2}$ – длина вектора X_1.

Ортонормируем второй столбец, положив

$$Z_2 = \alpha_{22} \cdot X_2 - \alpha_{21} \cdot Z_1.$$

Коэффициенты α_{22}, α_{21} определим из условий:

$$\langle Z_2, Z_1 \rangle = 0;$$

$$\langle Z_2, Z_2 \rangle = 1,$$

где $\langle Z_2, Z_1 \rangle$ – скалярное произведение векторов Z_2 и Z_1, которое будем вычислять по формуле

$$\langle Z_2, Z_1 \rangle = \sum_{i=1}^{N} Z_{2i} \cdot Z_{1i}.$$

Таким образом, в результате решения системы

$$\begin{cases} \alpha_{22} \langle X_2, Z_1 \rangle \alpha_{21} = 0, \\ \alpha_{22}^2 \langle X_2, X_2 \rangle - 2\alpha_{22}\alpha_{21} \langle X_2, Z_1 \rangle + \alpha_{21}^2 = 1 \end{cases}$$

имеем:

$$\alpha_{21} = \frac{\langle X_2, Z_1 \rangle}{\sqrt{\langle X_2, X_2 \rangle - \langle X_2, Z_1 \rangle^2}},$$

$$\alpha_{22} = \frac{1}{\sqrt{\langle X_2, X_2 \rangle - \langle X_2, Z_1 \rangle^2}}.$$

Для определения Z_3 получим соотношение вида

$$Z_3 = \alpha_{33} X_3 - \alpha_{31} Z_1 - \alpha_{32} Z_2,$$

где коэффициенты α_{33}, α_{31}, α_{32} определим таким образом, чтобы вектор Z_3 имел единичную длину и при этом был бы ортогонален векторам Z_1 и Z_2 т. е.

$$\langle Z_3, Z_3 \rangle = 1;$$

$$\langle Z_3, Z_1 \rangle = 0;$$

$$\langle Z_3, Z_2 \rangle = 0.$$

Из ортогональности следует

$$\begin{cases} \alpha_{31} = \alpha_{33} \langle X_3, Z_1 \rangle; \\ \alpha_{32} = \alpha_{33} \langle X_3, Z_2 \rangle. \end{cases}$$

Требование, чтобы вектор Z_3 имел единичную длину, приводит к необходимости решения уравнения

$$\alpha_{33}^2 \langle X_3, X_3 \rangle + \alpha_{31}^2 + \alpha_{32}^2 - 2\alpha_{33}\alpha_{31} \langle X_3, Z_2 \rangle - \\ -2\alpha_{33}\alpha_{32} \langle X_3, Z_2 \rangle = 1,$$

подставив в которое рассчитанные α_{31} и α_{32}, приходим к уравнению

$$\alpha_{33}^2 (\langle X_3, X_3 \rangle - \langle X_3, Z_1 \rangle - \langle X_3, Z_2 \rangle) = 1,$$

откуда следует, что

$$\alpha_{33} = \frac{1}{\sqrt{\langle X_3, X_3 \rangle - \langle X_3, Z_1 \rangle - \langle X_3, Z_2 \rangle}}.$$

В общем виде процесс ортогонализации описывается следующими рекуррентными соотношениями:

$$Z_k = \alpha_{kk} \cdot X_k - \sum_{i=1}^{k-1} \alpha_{ki} \cdot Z_i,$$

где

$$\alpha_{кк} = \frac{1}{\sqrt{\langle X_k, X_k \rangle - \sum_{i=1}^{k-1} \langle X_k, Z_i \rangle}},$$

$$\alpha_{ki} = \alpha_{kk} \langle X_k, Z_i \rangle,$$

$$i = \overline{1, k-1}.$$

Модель с минимальной погрешностью

Математическую модель будем строить в классе линейных моделей в следующем виде:

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \ldots + a_м z_м. \qquad (1)$$

где z_0 – фиктивная переменная, равная единице во всех опытах, коэффициенты a_i, $i = \overline{0, M}$ определим по методу наименьших квадратов (МНК).

Погрешность модели (1) представляет собой:

$$\varepsilon_0 = \max_{1 \le k \le N} \left(\left| y_k - \sum_{i=0}^{M} a_i z_{ik} \right| \right). \qquad (2)$$

Пусть $V \subseteq \Omega$ – перечень показателей, которые мы не будем включать в математическую модель типа (1) в качестве предикторов, и пусть $|V|$ – число элементов в этом перечне, тогда от-

клик y будет определяться с помощью показателей из множества Ω/V в следующем виде:

$$y = \sum_{x_i \in \Omega/V} a_i z_i \ .$$

Очевидно, что погрешность модели существенно будет зависеть от множества V и ее можно представить в виде

$$\varepsilon(\Omega/V) = \max_{1 \le k \le N} \left(\left| y_k - \sum_{\Omega/V} a_i z_{ik} \right| \right)$$

или

$$\varepsilon(\Omega/V) = \max_{1 \le k \le N} \left(\left| y_k - \sum_{i=0}^{M} a_i z_{ik} + \sum_{i \in V} a_i z_{ik} \right| \right).$$

Таким образом, оценка имеет вид

$$\varepsilon(\Omega/V) \le \varepsilon_0 + \max_{1 \le k \le N} \left(\left| \sum_{i \in V} a_i z_{ik} \right| \right)$$

или, обозначив через

$$\delta(V) = \max_{1 \le k \le N} \left(\left| \sum_{i \in V} a_i z_{ik} \right| \right),$$

можно утверждать, что

$$\varepsilon(\Omega/V) \le \varepsilon_0 + \delta(V).$$

Таким образом, возникает задача для заданного отклика y определить такой набор предикторов, чтобы погрешность $\delta(V)$ была бы как можно меньше, при этом число исключаемых из набора показателей $|V|$ было бы как можно больше.

В математическом плане сформулированная задача представляет собой задачу векторной оптимизации, и может быть представлена в виде

$$\begin{gathered} \delta(V) \to \min, \\ |V| \to \max, \\ V \subseteq \Omega . \end{gathered} \qquad (3)$$

Сформулируем основные свойства решения задачи (3).

Исходя из определения функции $\delta(V)$, следует

$$\delta(V_1 \cup V_2) \le \delta(V_1) + \delta(V_2),$$

т. е. она является полуаддитивной функцией множества.

Построим функцию

$$\overline{\delta}(V) = \sum_{i \in V} \delta(\{z_i\}),$$

для которой выполняется

$$\delta(V) \le \overline{\delta}(V).$$

Рассмотрим задачу векторной оптимизации для аддитивной вектор-функции множества V:

$$\begin{pmatrix} \overline{\delta}(V) \\ |V| \end{pmatrix} \to \min . \qquad (4)$$

Определим, что будем понимать под решением задачи (4) и сформулируем его основные свойства.

1. Множество элементов $V_* \subseteq \Omega$ будем называть эффективным, если любая его вариация приводит к увеличению или $\overline{\delta}(V)$, или $|V|$, или $\overline{\delta}(V)$ и $|V|$ одновременно.

2. Под решением задачи (3) будем понимать множество A, содержащее все эффективные наборы типа V_*, т. е. элементами множества A являются подмножества множества Ω, причем каждый из элементов представляет собой эффективное решение задачи (4).

Множество A называется множеством несравнимых вариантов по Парето. В рассматриваемой задаче это множество состоит из M элементов, представляющих собой структуры моделей, которые могут быть выбраны, исходя из заданной точности и числа предикторных переменных.

Пример. Рассмотрим задачу структурного моделирования для Приднепровской железной дороги, где в качестве исходной информации о деятельности дороги рассматриваются данные, приведенные в работе [5]. Деятельность предприятия будем определять по следующим показателям:

x_1 – грузооборот (млн т·км);

x_2 – пассажирооборот (млн пас-км);

x_3 – количество погруженных вагонов (тыс.);

x_4 – количество разгруженных вагонов (тыс.);

x_5 – производительность локомотивов (тыс. т·км брутто);

x_6 – вагонооборот (сутки).

В качестве отклика y рассмотрим x_1 – грузооборот. В соответствии с изложенной процедурой вычислим погрешности моделей $\overline{\delta}(\{x_i\})$, где $i = 2,6$.

Результаты сведем в табл. 1.

Таблица 1

x_i	$\bar{\delta}(\{x_i\})$	$\bar{\delta}$ %
x_2	4,698	2,4989
x_3	5,030	2,6755
x_4	8,820	9,4231
x_5	10,740	9,6931
x_6	5,010	2,6649

Заметим, что если включить в математическую модель в качестве предикторов все переменные, получим

$$x_1 = -63,459 - 0,346x_2 - 6,249x_3 + \\ +27,083x_4 + 0,113x_5 - 3,512x_6 .$$

Максимальная погрешность данной модели при этом составила $\varepsilon_0 = 2,2872\,\%$, отметим, что она оказалась меньше всех погрешностей $\bar{\delta}$ % из табл. 1.

Результаты решения задачи (4) представим в виде табл. 2.

Таблица 2

№ п. п	Ω/V	V	$\delta(V)$ %
1	$\{x_3, x_4, x_5, x_6\}$	$\{x_2\}$	2,4989
2	$\{x_3, x_4, x_5\}$	$\{x_2, x_6\}$	5,1638
3	$\{x_4, x_5\}$	$\{x_2, x_3, x_6\}$	7,8393
4	$\{x_5\}$	$\{x_2, x_3, x_4, x_6\}$	17,2624
5	$\{\ \}$	$\{x_2, x_3, x_4, x_5, x_6\}$	26,9555

Таким образом, если допустимая максимальная погрешность должна быть не больше 10 %, то структура математической модели с минимальным числом предикторов будет иметь вид

$$x_1 = f(x_4, x_5),$$

в этом случае математическую модель можно представить в следующем виде

$$x_1 = a_0 + a_4 x_4 + a_5 x_5 , \qquad (5)$$

где параметры модели a_i – определяются по методу наименьших квадратов и равны:

$$a_0 = -73,570;$$

$$a_4 = 21,213;$$

$$a_5 = 0,097,$$

при этом максимальная погрешность составляет 2,2 %.

Таким образом, грузооборот (x_1) определяется количеством разгруженных вагонов (x_4) (тыс. сут.), и производительностью локомотива (x_5) (тыс. т·км брутто/сут.).

На рис. 1 представлены наблюдаемые и рассчитанные по модели (5) значения грузооборота (x_1).

Для сравнения приведем математическую модель (6), когда в качестве предикторов взяты все показатели x_2, x_3, x_4, x_5, x_6.

$$x_1 = -63,459 - 0,346x_2 - 6,250x_3 + \\ +27,083x_4 + 0,113x_5 - 3,512x_6 . \qquad (6)$$

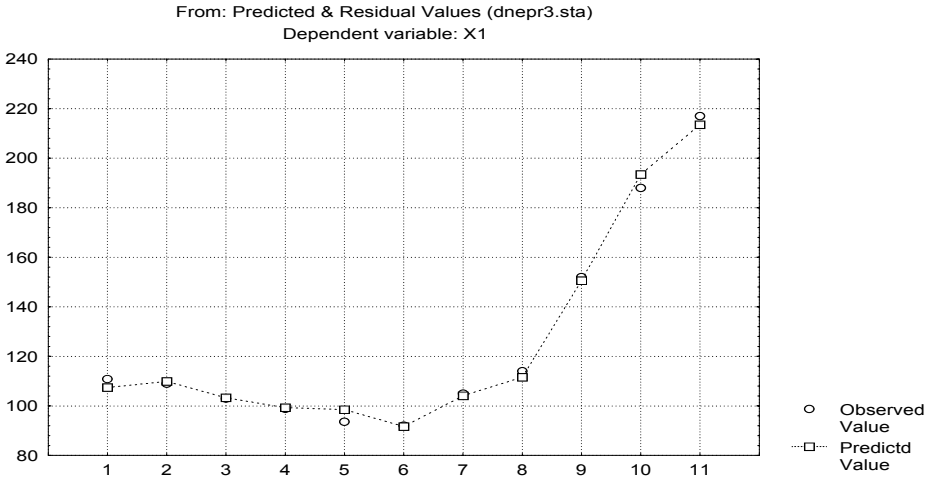

Рис. 1. Сравнение наблюдаемых и рассчитанных по модели (5) значений грузооборота

На рис. 2 представлены наблюдаемые и рассчитанные по модели (6) значения грузооборота в зависимости от пассажирооборота, количества погруженных и разгруженных вагонов, производительности локомотива и вагонооборота.

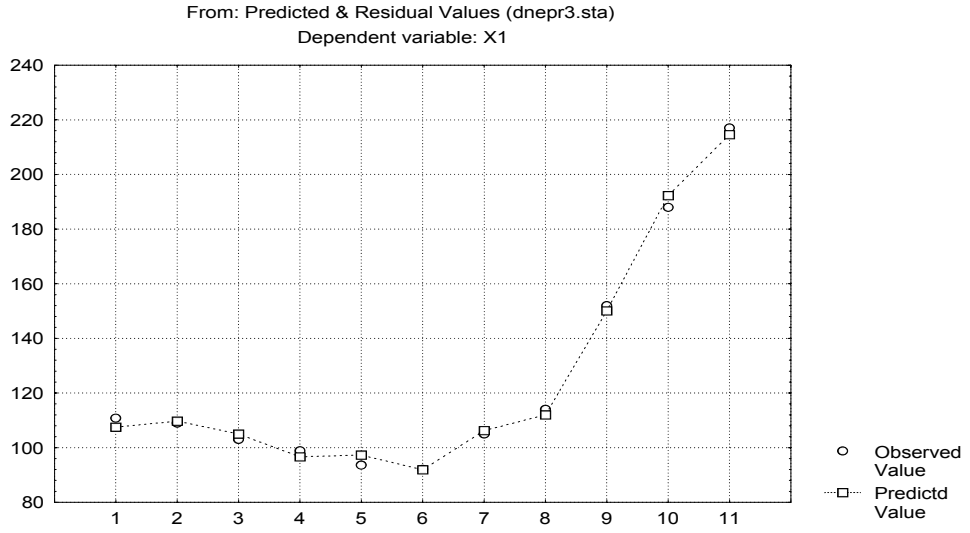

Рис. 2

Выводы

1. Предложена методика выбора предикторных переменных с определенным порядком применения регрессионного анализа.

2. Получена возможность выбора предикторных переменных по заданной точности математической модели.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 1. – М.: Финансы и статистика, 1986. – 366 с.
2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 2. – М.: Финансы и статистика, 1987. – 352 с.
3. Босов А. А., Мухина Н. А. Основные задачи моделирования по экспериментальным данным // Питання прикладної математики та математичного моделювання: Зб. наук. праць ДДУ. – Д., 1999. – С. 7–12.
4. Боровиков В. П., Боровиков И. П. STATISTICA, Статистический анализ и обработка данных в среде WINDOWS. – М.: Филин.
5. Пасечкін В. І. Аналіз динаміки показників залізниць України (за результатами моніторингу за період 1991–2001 рр.). // Залізничний транспорт України № 5, 2002. – С. 2–6.

Поступила в редколлегию 22.10.03.